DOCUMENT RESUME

ED 345 513 FL 020 178

AUTHOR Alderson, J. Charles; Wall, Dianne
TITLE Does Washback Exist?
PUB DATE Feb 92
NOTE 23p.; Paper presented at a Symposium on the Educational and Social Impacts of Language Tests, Language Testing Research Colloquium (February 1992). For a related document, see FL 020 177.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Techniques; Educational Environment; Educational Research; Educational Theories; Foreign Countries; Language Research; *Language Tests; *Learning Processes; Literature Reviews; Research Needs; Second Language Instruction; *Second Languages; *Testing
IDENTIFIERS Nepal; Netherlands; *Teaching to the Test; Turkey

ABSTRACT

The concept of washback, or backwash, defined as the influence of testing on instruction, is discussed with relation to second language teaching and testing. While the literature of second language testing suggests that tests are commonly considered to be powerful determiners of what happens in the classroom, the concept of washback is not well defined. The first part of the discussion focuses on the concept, including several different interpretations of the phenomenon. It is found to be a far more complex topic than suggested by the basic washback hypothesis, which is also discussed and outlined. The literature on education in general is then reviewed for additional information on the issues involved. Very little research was found that directly related to the subject, but several studies are highlighted. Following this, empirical research on language testing is consulted for further insight. Studies in Turkey, the Netherlands, and Nepal are discussed. Finally, areas for additional research are proposed, including further definition of washback, motivation and performance, the role of educational setting, research methodology, learner perceptions, and explanatory factors. A 39-item bibliography is appended. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Charles Alderson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Does washback exist?

J Charles Alderson and Dianne Wall
Lancaster University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED345513

Paper presented in the symposium on "The Educational and Social Impacts of Language Tests",
Language Testing Research Colloquium, February 1992

Introduction

The notion of washback, or backwash - the influence of tests on teaching - is commonplace in the educational and applied linguistic literature. There has developed within British applied linguistics a tendency to use the term "washback" to label the phenomenon of the influence of testing on teaching, although the older term "backwash" is still prevalent in the educational literature. Since discussions of this phenomenon within British language testing for the past twenty years have tended to use the term "washback" rather than "backwash", we continue the tradition, and use this term. Those who prefer the term "backwash", for whatever reason, are invited to consider "washback" to be a simple translation equivalent. We are ourselves persuaded that the difference in terminology has no semantic or pragmatic significance whatsoever.

It is commonly asserted that tests have influence: that is, that tests affect teachers and learners and thereby affect teaching and learning. (See, for example, Wiseman, 1961; Davies, 1968; Kellaghan, 1982; Alderson, 1986; Morrow, 1986; Pearson, 1988, Hughes, 1989; Khaniya, 1990a and b). For example, Pearson (1988), says:

> "It is generally accepted that public examinations influence the attitudes, behaviour, and motivation of teachers, learners and parents."

Ebel (1979:23, quoted in Khaniya, 1990b, claims that it is common practice for students to work harder when they know that they are approaching exams than when they do not.

In describing the effect of examinations, Wong (1969) writes:

> "The examination dictates the activities in schools. Syllabuses .. are issued by examination syndicates and central authorities.
> Interpretation of the syllabus is carried out chiefly by reference to past examination papers which.. tend to carry questions similar in type and content year after year."
> (Wong, 1969, p363 quoted in Khaniya (1990b).

FL 020179

BEST COPY AVAILABLE

Morris (1972:75) considers examinations necessary to ensure that the curriculum is put into effect, and Wiseman (1961:64, quoted in Khaniya (1990b), says that "good examinations are useful and desirable: without them education would be poorer and much less effective".

Pilliner (1973:4) claims that the most important requirement of a good test is that it should be educationally beneficial, thus taking washback for granted.

Alderson (1986: 104) discusses the "potentially powerful influence of tests", and argues for innovations in the language curriculum through innovations in language testing.

Khaniya (1990b: 22) asserts that washback is an inherent attribute of an examination:

"Since an examination is used as an achievement test, asking students to take an exam entails teaching and preparing for it....Whatever is done all along the way (sic) of examination preparation is the 'washback' effect of the examination. This effect can influence the teaching and learning methods employed from beginning to end of a course if examinations require students to cover all what (sic) is entailed in the course objectives. But if an exam does not require the students to work for the whole year, the whole preparation will rest on the last couple of weeks/ months before the examination".

Khaniya also asserts that "an exam defines for the students the content and performance objectives of the course" (p 26).

Pearson (1988) goes even further and claims not only that good tests will encourage the use of "beneficial teaching-learning processes", but also that they

"will be more or less directly usable as teaching learning activities. Similarly, good teaching-learning tasks will be more or less directly usable for testing purposes, even though practical or financial constraints limit the possibilities."

None of these assertions appear to us to be in any way unusual. Washback is often introduced on language testing courses as a powerful concept that all test designers need to pay attention to, and which most classroom teachers are all too aware of. Swain (1985) discusses the importance of the influence of the test on teaching, and recommends that test developers "bias for best" and "work for washback". (Although quite how test designers are to take account of even potential washback, much less actually experienced washback, is to our knowledge not discussed.) Davies (1985) asks whether tests necessarily follow the curriculum, and suggests that perhaps tests ought to be leading and influencing curricula.

Some writers have even gone so far as to suggest that a test's

validity should be measured by the degree to which it has had a beneficial influence on teaching. Keith Morrow (1986) coined the term "washback validity" to denote the quality of the relationship between a test and associated teaching. The notion presumably means something like: "this test is valid when it has good washback"; and conversely, "this test is invalid when it has negative washback". He says (p6):

"The first validity criterion that I would .. put forward for (these examinations) would be a measure of how far the intended washback effect was actually being met in practice".

He admits, however: "I am not sure at this stage how it could be measured", although he then goes on to claim:

"In essence an examination of washback validity would take testing researchers into the classroom in order to observe the effect of their tests in action".

He cites Wilkins, Widdowson and others as asserting that direct tests of language performance will be "most beneficial in terms of washback effect", and argues that communicative tests like the former RSA CUEFL should have a "powerful and positive washback effect into the classroom" (p6). This sentiment was echoed recently in the general educational literature,by Frederiksen and Collins (1989), whose criteria for a valid test include the degree of directness of assessment of cognitive skills, and the degree of subjective judgement that is required to assign a score. They consider that valid tests will involve the subjective, direct assessment of higher-order cognitive skills

Indeed, Frederiksen and Collins introduce a concept similar to "washback validity", with a different name: the term "systemic validity", which they define as follows:

"A systematically valid test is one that induces in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time" (p27)

However, to our knowledge, this form of validity has never been demonstrated, or indeed investigated, nor have proposals been made as to how it could be established empirically rather than asserted. Moreover, it is not at all clear that if a test does not have the desired washback this is necessarily due to a lack of validity of the test, as Morrow and others simplistically imply. It is surely conceivable that other forces exist within society, education and schools that might prevent washback from occurring, or that might affect the nature of washback despite the "communicative" quality of a test. This can then hardly be attributed to a problem with the test. Whereas validity is a property of a test, in relation to

its use, we will argue that washback, if it exists - which has yet to be established - is likely to be a complex phenomenon which cannot be related directly to a test's validity.

It seems to us to be important to investigate the nature of washback first, and the conditions under which it operates - what affects it, how teachers and students prepare for tests, what the nature is of the relationship between a test and teaching. Only once we understand more about the nature of washback - once we are able to describe what actually happens, will we be in a position to explore why these things happen - what "causes" these effects. And only after we have established causal relationships to washback will we be in a position to explore whether we are justified in relating washback to a test's validity . Thus, talk of washback or systemic validity (already fashionable in some circles, see Khaniya, 1990a and Weir, 1998) is at best premature, and at worst ill-conceived.

In summary, the term washback is common in the language teaching and testing literature and tests are held to be powerful determiners of what happens in classrooms. However, the concept is not well defined in the literature, and we believe that it is important to be more precise about what washback might be before we can investigate its nature. In addition, we need to distinguish between description and explanation: it is important to establish what washback actually looks like in classrooms and elsewhere, before we can hope to explain it. It is probably also important to distinguish between influence / impact and washback. Thus, it is at least conceivable that it might be useful to talk about the influence of a test on teachers' attitudes - to the tests themselves, to the syllabus and to their teaching. Similarly, pupils and parents might have attitudes to, opinions about, tests which influence their behaviours. However, this is not the same as, although conceivably related to, the influence of the test on teaching and learning, ie what actually happens in classrooms that can (or not) be attributed to The Test. Whilst the topic of this paper is washback, it is important to remember that test impact is a wider issue, and likely to be important for an understanding of what actually happens in classes.

This paper is in several parts: first we speculate upon some possible interpretations of the phenomenon. Then we refer to the general educational literature for enlightenment on these issues since the applied linguistic literature appears to take the phenomenon of washback for granted rather than to question it. Next, we look at what empirical research exists in the language testing field for the insights it can offer. Finally, we briefy present a series of proposals for further research.

Exploring the concept of washback

The belief in washback refered to above is most commonly asserted with respect to negative washback, namely the

supposed negative or undesirable effect on teaching/learning of a particular (and, by inference if not direct statement, "poor") test. In this case, "poor" usually means "something that the teacher or learner does not wish to teach or learn".

It has often been observed that washback need not be negative: the term 'washback' implies influence, of any sort. If the test is 'poor', then the washback is felt to be negative. Logically, if the test is 'good', then its influence could be positive - if the Washback Hypothesis (WH) holds, then good tests should have good effects (as yet undefined).

If we consider these beliefs briefly, we can see that other possibilities also hold. The Washback Hypothesis seems to assume that teachers and learners do things <u>they would not necessarily otherwise do because of the test.</u> Hence the notion of influence. But this also implies that a 'poor' test could conceivably have a 'good' effect if it made teachers and learners do 'good' things they would not otherwise do: for example, prepare lessons more thoroughly, do their homework, take the subject being tested more seriously and so on. And indeed, it is relatively commonplace to note that teachers often use tests to get their students to do things they would not otherwise do: to pay attention to the lesson, to prepare more thoroughly, to learn by heart, and so on. To the extent that these activities are in some sense desirable - hard-work is presumably more 'desirable' than no work at all, extrinsic motivation might be better than no motivation at all - then any test, good or bad, can be said to be having beneficial washback if it increases such activity or motivation.

Alternatively, one might wish to consider the case where any test has negative effects. The most obvious candidate for this is anxiety in the learner brought about by having to take a test of whatever nature, and, if not anxiety, then at least concern in teachers, if they believe that some consequence will follow on poor performance by the pupils. The argument would go like this: any learner who is obliged to do something under pressure will perform abnormally, and may therefore experience anxiety. Thus pressure produces abnormal performance, the fear of which produces anxiety. In addition, the fear of the consequences of particular performances produces anxiety which will influence performance. Similarly for teachers, the fear of poor results, and the associated guilt, shame, or embarrassment, might lead to the desire for their pupils to achieve high scores in whatever way seems possible. This might lead to "teaching to the test", with an undesirable narrowing of the curriculum.

We may also wish to consider the case where the test <u>reinforces</u> some behaviour or attitude rather than bringing about an otherwise unlikely behaviour. Thus students may already work hard, and a test may simply motivate them to work harder. A learner may constantly self-evaluate against internal or external criteria, and the test may provide very useful additional criteria against which to compare oneself.

Thus the relationship between a test and its impact, positive or negative, might be less simple than at first sight appears to be the case. The quality of the washback might be independent of the quality of the test.

The question arises as to whether 'washback' is the same as 'influence' or whether the term refers solely to some sorts of influence and not others? Thus, we might not want to call anxiety caused by having to take a test 'washback', but syllabus or textbook design specifically based on a test (eg the Longman series of textbooks intended to prepare students for the Cambridge First Certificate in English examination) we might indeed want to call washback.

Even if we were to use the term 'washback' to refer to the test's effect on textbook design, we would probably need to distinguish between pedagogic material which is directly related to a specific test (in content or method, etc, for example, TOEFL preparation courses) and material which is intended to help students get ready for an exam in some more general way, for example Study Skills courses which claim they give students skills relevant to taking an EAP test like the IELTS. Given these complexities, we may wish to restrict the use of the term washback to classroom behaviours of teachers and learners rather than to the nature of print and other pedagogic material. It is not clear from the literature, however, that writers do indeed so intend the term to be interpreted.

Another aspect of the notion of washback that needs examination is its deterministic nature: how directly, according to the WH, do tests bring about change in teaching/ learning? A naive deterministic view (which is often implicit in the complaints about TOEFL, for example, or even in the claim that tests can be used as "levers for change") would assume that the fact of a test having a set of qualities is sufficient in itself, by virtue of the nature of the importance and influence of tests in most societies, to bring about change. However, what we know about change is that it is not quite so simple: what influences how, when, etc teachers and learners change their behaviours/ beliefs, etc is certainly complex.

Discussion of washback in the literature tends to assume that the existence of a test brings about some change in motivation and thus in behaviour. In fact, the relationship between motivation and performance is a very complex matter beyond the scope of this paper to explore. However, a thorough study of washback must surely take account of research findings in this area. In fact, there appear to be conflicting results, as Fransson's brief review indicates. He points out that up to an optimal point, an increase in level of motivation is accompanied by an increase in learning. However, beyond that point an increase in motivation seems to have negative effects and performance declines (the so-called Yerkes-Dodson Law, Fransson, 1984). The position of the optimal point, Fransson

suggests, depends upon the difficulty of the task. However, it may well also relate to the consequences of the task (in our case the test), as well as to other factors within the performer such as that person's need for achievement (nAch). As McDonough (1981) points out, the strength of nAch itself may be the result of two opposed tendencies: the motivation to success and the motivation to avoid failure. Each of these two tendencies can be thought of as composed of three factors (McDonough, 1981, p146):

1) the person's expectations of success (or failure)
2) the value of the task as an incentive
3) the person's orientation toward success or toward avoidance of failure .

As if this were not sufficiently complicated, McDonough (op cit) goes on to review a further theoretical position, that of attribution theory, which describes motivated behaviour in terms of "the causes to which the individuals attribute or ascribe their own and other people's performance: their own ability, effort, intention or other's ability effort and intention, luck and so on".

It may however be that the key factor is not motivation (or extrinsic motivation, as Biggs and others have pointed out - see Fransson, 1984) but anxiety, both state anxiety - the condition in which you find yourself when performing a task - and trait anxiety - one's habitual response to stress. Furthermore, it may be important to distinguish two sorts of anxiety: debilitating and facilitating. Which of these is aroused in a particular learner or teacher may depend on personality factors, (eg extroversion/ introversion, need for achievement, fear of failure, and so on) as well as the consequences (and the learners' perception of those consequences) of particular performances.

What this brief excursion into motivation and anxiety is intended to illustrate is the extreme complexity of the topic, and the contrasting naivety of the Washback Hypothesis: clearly those asserting the existence of washback need to take more account of research findings and resulting theoretical positions in related fields.

The point we are making is that the Washback Hypothesis is unduly simplistic and makes too many untested assumptions about how people are influenced. This applies as much to negative washback (eg the assertion that TOEFL forces people to do certain things) as it does to positive washback. However, it will be important when empirically examining washback to look at both negative and positive situations, to see how comparable they are.

The Washback Hypothesis (es)

It might help to clarify our thinking a little if we attempt

to state the Washback Hypothesis explicitly. From a reading of the literature on language testing generally, and from our experience of talking to teachers about their teaching and testing, it is possible to develop different hypotheses, from the most general and vague to somewhat more refined hypotheses. which take account of different factors. It might help our thinking if we try to separate out the factors, as below.

Some Possible Washback Hypotheses (WHs)

1) A test will influence teaching.

This is the WH at its most general. However, by implication:

2) A test will influence learning
Since it is possible to separate the content of teaching from the methodology:

3) A test will influence how teachers teach

and

4) A test will influence what teachers teach

and therefore by extension from 2) above:

5) A test will influence what learners learn

and

6) A test will influence how learners learn

However, perhaps we need to be somewhat more precise about teaching and learning, whence:

7) A test will influence the rate and sequence of learning

and

8) A test will influence the rate and sequence of teaching

and the associated:

9) A test will influence the degree and depth of learning

and

10) A test will influence the degree and depth of teaching

If washback relates to attitudes as well as to behaviours, then:

11) A test will influence attitudes to the content, method etc of learning/ teaching.

In the above, no consideration has been given to the nature of the test, or the uses to which scores will be put. It seems not unreasonable to hypothesise:

12) Tests that have important consequences will have washback.

and conversely

13) Tests that do not have important consequences will have no washback.

It may be the case that:

14) Tests will have washback on all learners and teachers.

However, given what we know about differences among people, it is surely likely that:

15) Tests will have washback effects for some learners and some teachers, but not for others.

Thus one variable is teaching; learning is a related but in principle separate variable. A further set of variables relates to the content of the test and the content of the teaching/ learning. Another is the nature of the learning/ teaching: its rate, sequence, degree, depth, methodology. The importance of the consequences of performance on the test is another variable that needs to be considered, as is the complex of variables operating within individuals.

Clearly, we are complexifying what was initially a simple assumption. Is this justified? Is washback a concept to be taken seriously, or simply a metaphor which is useful in that it encourages us to explore the role of tests in learning and the relationship between teaching and testing? We are not sure at present, but we suspect that if it is a metaphor, it needs to be articulated somewhat more precisely if it is to throw light on teaching and testing, or indeed on the nature of innovation and change. And if it is a concept to be taken seriously, then we need to examine it critically, and see what evidence there might be that could help us in this examination.

Hence, we need in either case to identify cases where washback might be thought to have occurred, and to see what, how and why it did or did not occur.

Research into washback

The general educational literature

Surprisingly little empirical research has been conducted into the nature or indeed existence of washback in education in

general, much less in language education. Insofar as there can be said to be "classic" studies in a field where there are few studies, the "classic" study - the study that is perhaps best known - is that by Kellaghan, Madaus and Airasian, into "The Effects of Standardized Testing", published in 1982, but as the authors themselves confess, "conceived in the 1960s and planned and executed in the 1970s".

This joint Irish-American study examined the impact on Irish schools of introducing standardized tests. Ireland was chosen because, unlike the USA and elsewhere, there was no tradition of standardized testing in existence. It was thus possible to introduce such tests selectively into experimental schools and contrast outcomes and attitudes with control groups of various sorts.

The study, longitudinal in design, took place from 1974 to 1977. In the experimental group, norm-referenced standardized tests of general ability (the Otis-Lennon IQ test) and achievement (in mathematics, Irish and English) were given to pupils in grades 2 to 6, and resulting norm-referenced information was given to teachers. In one control group of schools, no testing was carried out. In the other, the tests were administered, but no results were given to teachers. The study investigated school, teacher, pupil and parent level effects. At school level, the study looked at the effect of tests on school organization and practice, and on school achievement. At teacher level, it looked at teachers' attitudes, beliefs and behaviour in relation to standardised tests, teachers' reactions to testing and teachers' perceptions of the usefulness and reported use of test information. The researchers looked at pupils' perceptions of and reactions to standardized tests, their perceptions of factors that affect scholastic progress and getting along well in class, their ratings of their scholastic behaviour and abilities, and their self-concepts, and they also investigated, with respect to parents, their familiarity with changes in evaluation, the communication between school and parents, parents' perception of their children's school progress and their knowledge about and attitudes toward standardized testing. In addition, the study investigated expectancy effects and the role of test information in teacher expectations and perceptions of pupils. The study asked a wide range of questions, such as "Does the availability of standardized test information affect teachers' perceptions of pupils? Does it affect levels of student achievement? How much weight do teachers give to standardized test information relative to other types of evidence in making educational decisions? Do teachers perceive certain types of tests as biased against certain groups of pupils? Does the content of standardized tests affect the nature and emphasis of classroom instruction?" and many other questions. It will be evident from this that "washback" defined as the impact of tests on classrooms was not the only matter under investigation. Parent perceptions, school administrative systems, the sorts of information used to make educational decisions, and above all

the expressed opinions and perceptions of tests by various participants were at least equally focussed upon.

The results showed very little effect of standardized tests on school organizational or assessment practices - they tended to be used to support rather than to disrupt existing practices. In fact, test scores tended to confirm teachers' ratings of pupils' ability and achievement rather than the reverse. Nothing in the findings supported the belief that taking tests regularly leads to increased performance on tests: a comparison of experimental and control groups' performances on ability and achievement tests at the end of the four-year period revealed complex results that could be interpreted in a variety of ways, but which did not provide evidence of a simple effect of practice or familiarity with testing or the provision of test results (page 61).

The reactions of teachers and pupils to the test programme were very positive - the tests were perceived as fair and accurate, as providing stimulation rather than anxiety, yet pupils took the test seriously and a large majority reported enjoyment of the experience rather than fear. Teachers who had received tests and test information were more positive about tests and their value than were the control groups. Parents showed very little impact of the experimental tests: they were largely unaware of the existence of such tests, although their attitudes to testing in general were very positive. With respect to pupils, there was no evidence that the provision of test information had negative effects on pupils' self concepts, or their self-assessments. There was evidence of an expectancy effect, such that teachers in receipt of test information about their pupils rated their pupils in line with that test information. However, expectancy effects were at work regardless of whether teachers had test information, or other information or expectations about their pupils. On the whole, the provison of test information appeared to work to the pupils' advantage: provided with discrepant test information, teachers gave pupils the benefit of the doubt.

However, one criticism of the value of these results is that the situation was artificial: the tests that were introduced for the sake of the experiment had no currency or consequence within the Irish educational system. Pupils were not passed or failed, they were not denied entry to secondary or tertiary level education, promoted to the next class or held back as a result. The results of the tests were thus of no significance, and it is not surprising that relatively little negative impact was perceived under such circumstances. The Washback Hypothesis, on the other hand, presumably applies to tests and examinations that are used regularly within the curriculum. All schools are presumably affected by them, and not just some, and such tests can be presumed or perceived to have educational consequences. To such settings, the Kellaghan et al study has little of direct relevance.

A second drawback with the Kellaghan et al study from the

point of view of this paper is that the dependent variables were either teacher ratings of pupils, test scores, or questionnaire responses. Very little usable information was gathered independently on what happened in the experimental and control group classrooms, and we are obliged to rely upon teacher and pupil accounts of practices. Although the study is very valuable for what it reveals with respect to a variety of aspects of test impact, particularly in relation to perceptions, observational evidence of test impact on classroom teaching/learning is minimal.

Indeed this lack of evidence from classrooms is a characteristic of virtually all writings about the influence of tests on teaching. See, for example, Paris, Lawton, Turner and Roth, 1991, Haladyna, Nolen and Haas, 1991, or Frederiksen, 1984, all of whom use anecdote, assertion or interviews and surveys of what teachers and pupils say they do, rather than direct observation.

An exception is Smith (1991) who reports on two qualitative studies which investigated the effect of tests on teachers and classrooms. Data from interviews revealed that the publication of test results induced feelings of fear, guilt, shame, embarrassment and anger in teachers, and the determination to do what was necessary to avoid such feelings in the future. Teachers reportedly believed that test scores were used against them, despite the perceived invalidity of the scores, and they also believed that testing had severe emotional impact on young children (less so on older pupils). From classroom observation it was concluded that testing programmes substantially reduced the time available for instruction and narrow the curriculum and modes of instruction:

"What we saw in one school's sixth grade was a transition, as the school year progressed toward ITBS testing in April, from laboratory, hands-on instruction in science several days a week, to less frequent science out of textbooks (choral reading from the text and answering comprehension and vocabulary questions on worksheets), to no science instruction at all in the weeks before the test, to either no science at all or science for entertainment value during the ITBS recovery phase, to science instruction precisely tailored to the questions in the district criterion-referenced tests, to no science at all. The same group devoted about 40 minutes each day to writing projects in the fall, but the class wrote no more after January, after which they spent the time on worksheets covering grammar, capitalization, punctuation and usage. Writing instruction returned in late May, when the pupils again began producing poetry, stories, reports on projects for the short time remaining in the school year. Social studies and health instruction disappeared altogether." (Smith, 1991, p10)

Interestingly, however, Smith reports that there were two different reactions to this "narrowing of the curriculum". One was accommodation by teachers, who discarded what was not

going to be tested, and taught towards the test, but the other was one of resistance, exemplified by this quotation from one teacher: "I know what's on the test, but I feel that these children should keep up with current events and trace the history behind what's happening now, so we're going to spend March doing that. I guess I'm saying that the test scores are going to be up for grabs" (Smith, 1991, p 10). This suggests that the wshback phenomenon is not quite as simple as is at times made out. We need many more studies like those Smith reports before we can claim we understand the nature and mechanisms of washback.

Language education

There is remarkably little in the specific field of language education that can be said to have investigated and established what washback is and how it works. Much assertion exists, for example, the debate about the influence of the introduction of multiple choice tests in Ethiopia - see Forbes (1973) and Madsen's reply (Madsen, 1976). Forbes attacks objective test methods, and makes claims about what happens in classrooms:

"Gone are the happy days in which a teacher could spend a whole period on his (sic) favourite poem, 'The Solitary Reaper' if he wanted to. He may not even spend time on Belloc's 'Tarantella', even though it is in the prescribed textbook written by some of the university 'English language specialists'...So it's eyes up to the sentences on the blackboard - sentence patterns for tenses, for quantifiers, for modals, for relative clauses. Which is right and which is wrong? Write them down to remember them, perhaps, but don't write anything else. That's waste of time (sic). We are back to "The pen of my aunt is in the garden" and "The postillion was struck by lightning" with a vengeance." (p135)

Sadly (or perhaps not surprisingly) Forbes provides no evidence to support his emotive claims, nor does Madsen in his reply. Even when justifying the introduction of objective tests in terms of how teachers were preparing students for the old examinations that the objective tests replaced, Madsen has to resort to impressions:

"Teachers <u>appeared</u> to be short-changing their students in the classrooms. English teachers in the upper grades in particular <u>seemed</u> to be spending virtually all their time on examination techniques rather than on the English fundamentals so badly needed" (p136, our underlining)

Similarly, when describing the effect of the objective test, unsupported claims are made: "Teachers in the upper grades were <u>inclined to</u> model instruction on the now sacrosanct objective examination...the backwash effect on the schools became just as devastating as that produced by the earlier precis-essay examination" (page 138, our underlining).

The only projects that are known to the current authors in language education that have systematically investigated the phenomenon relate to the Netherlands (Wesdorp, 1982), Sri Lanka (Alderson et al, 1987, Pearson, 1988, Alderson and Wall 1990, 1991 and Wall, 1991), Nepal (Khaniya, 1990a and 1990b), Turkey (Hughes, 1988) and China (Li Xiaoju, 1989).

Information, published or unpublished, may exist with respect to other countries, and we would be very interested to hear of such information. It is, in our view, nevertheless noteworthy that we have failed to uncover more empirical studies, given the firmness with which a belief in washback is held in language teaching circles. What follows is an account of those studies that have been identified to date.

The Netherlands

Wesdorp (1982) gives an unpublished account of research into the validity of objections to the introduction of multiple-choice tests into the assessment of mother tongue and foreign language education. The research found that most of the objections, which assumed washback effects, were not justified. It was, for example, assumed that skills that could not be tested by multiple-choice would not be taught in primary schools, but a comparison of essays written before the introduction of mcq writing tests, and twelve years after that introduction, found no differences in quality. An investigation of differences in teacher activities in schools with and without a mcq final tests failed to show any clear differences. No evidence was forthcoming of an increased use of mcq in language teaching, nor of any change in student study habits as a result of mcq tests in English (interestingly, there was evidence of a relationship between study habits and test formats for subjects other than English). In short, empirical investigation revealed much less washback effect than had been feared.

Turkey

Hughes (1988) describes a project at Bogazici University, Istanbul, where innovations were made in test design with a view to bringing about change in the curriculum. Prior to the start of the project, students were entering main stream academic studies after a year's preparation at the Foreign Language School (FLS) with very low levels of English proficiency. Undergraduate teaching staff complained of the level of English of incoming students. Test evidence showed that fewer than 50% of the students completing studies at the FLS achieved a minimally acceptable score on the Michigan test, yet 99% of students were admitted into their main subject areas. As a result of this poor English performance, it was decided that a new proficiency test should be designed which would be the sole criterion for determining whether students should proceed to undergraduate studies. A new test was designed whose content reflected the sort of uses of

English that might be expected in an English-medium university like Bogazici. The immediate reported effect was consternation at the standards set by the new test, together with a realisation that if changes at FLS did not occur, then many students would fail the test. As a result, teaching syllabuses were changed, new textbooks introduced, the number of contact hours increased and

> "for the first time, at least for some years, FLS teachers were compelled, by the test, to consider seriously just how to provide their students with training appropriate for the tasks that would face them at the end of the course." (Hughes, 1989, p 144)

Only 50% of the first batch of students managed to pass the new proficiency test, although this rose to 86% by the end of an intensive summer school. The evidence was that standards of English had indeed risen, since at the end of the first year in which the new proficiency test was introduced, between 72% and 83% of students achieved the minimum acceptable Michigan score, and a survey of academic staff showed that the English proficiency of students entering mainstream studies was perceived to be "very, very much better" than their predecessors. Hughes (1988) claims that this state of affairs came about because of the beneficial washback effect of the test:

> "Teaching for the test (which may be regarded as inevitable) became teaching towards the proper objectives of the course" (page 145) (since the test was based directly on the English language needs of undergraduate students: our explanation)

and goes on to argue that

> "potential backwash effect should join validity and reliability in the balance against practicality. If this were done, one might find that there were fewer conflicts between teaching and testing than appear to exist today."

Hughes seems to demonstrate that tests can indeed impact on the language curriculum, especially if their consequences are important, as in the case of the Bogazici proficiency test. Certainly, changes in the syllabus, textbooks and possibly in the teaching in the FLS are reported to have occurred, and this appears to have been due to the proposed introduction of the new test. It also seems to be the case that something associated with these changes brought about improved levels of proficiency in English, although what that something is, is unclear. It is at least conceivable that <u>the mere threat</u> that students might actually be failed on a proficiency test of whatever nature led to students and teachers working harder, but not necessarily in the "right" (ie intended) direction. Curiously, although the new proficiency test was quite unlike the Michigan test in content and method, presumed preparation for the new test resulted in increased proficiency defined

very differently. Hughes unfortunately does not address this issue. Nor do we know what actually changed in classrooms; in short we do not know what washback effect the test produced, nor how it produced it. Thus, although increases in English proficiency were established, the origin of these is uncertain. Nevertheless, Hughes presents evidence for possible washback that suggests that it would indeed be worth investigating further how tests can bring about change.

Nepal

In Nepal the SLC (School Leaving Certificate) is an extremely important hurdle to tertiary education and good employment as well as social status. Khaniya (1990b) describes the existence of published cribs for the exam with exotic titles like "Gautam Super Lucky SLC Guess Paper" and "Guess Paper with High Surety for SLC". In fact, the SLC as described requires students to memorise texts and answers to questions, since many of the test questions and texts are taken directly from the textbooks, and are actually not answerable without reference to the textbook (or a memorised version of it). In such circumstances, where memorisation is essential to successful (or even unsuccessful) performance, it is perhaps not at all surprising that exam coaching occurs, and visible signs of this, like the publications mentioned, are clear evidence of washback in some form. However, even here we have no description of how teachers actually teach to the exam, what and how students learn, and so on. In fact, Khaniya reports very high failure rates on the SLC (90%), which must mean that if cramming is necessary for the exam, and if, as he asserts, cramming is rife, it must either be very inefficient, or the exam must be more unpredictable than the writers of cribs and cramming courses admit, or than Khaniya himself describes. What we clearly need is a description of what this claimed washback actually looks like and how and when it is successful.

Interestingly, Khaniya's results show that of the four types of schools he investigates, Type A schools (teaching in English medium, reportedly doing no coaching for the exam, and indeed not holding the exam in much regard) get the highest scores on the SLC! Whilst the Type D schools, which are asserted to engage in the most examination preparation and where therefore the washback effect is assumed to be highest, gain the lowest scores on SLC. In other words, if teachers teach English well and don't allow washback, their pupils will do well. If teachers teach English minimally and engage in exam coaching, then their children will do poorly!

Khaniya's attempt at investigating washback is, in fact, indirect. What he did was to design a new exam, on recent "communicative" lines, designed to be relevant to the use of English in tertiary education, and intended to engineer beneficial washback, and then compared his new exam with the SLC. (The new exam included two reading passages and multiple

choice and one-word-answer questions, a random (dr=6) cloze, a note-taking test and two compositions (a letter and an essay, but no tests of listening and speaking, for practical reasons.) He gave his new exam to students at the beginning and then at the end of Grade 10 (when students are preparing for the SLC). Since they did not on the average improve in their performance on the new exam, he claims that this is because students are cramming for the SLC, which cramming does not "teach English". Unfortunately Khaniya was not able to administer the SLC at the beginning and end of the year as well: it could be that the ability to take SLC did not improve either.

He was able to compare performance on the SLC at the end of the year with performance on the new exam (taken at the beginning of the year) and showed that for all but Type A schools, the SLC scores were higher than the new exam. Type A school students did better on the New Exam than the SLC. He claims, therefore, that the students are learning SLC-ese, not English. However, this assumes that in some sense the tests should be equivalent in difficulty, which he does not establish. In other words, one interpretation of the results is that students might be finding the New Exam more difficult because it is! (Although Type A school results suggest that this is not the case.) The result could in any case have more to do with method effect than anything else: cloze, note-taking and letter writing were particularly difficult for the students, and they might be expected to be most unfamiliar. Of course, the notion of familiarity with test method - method effect - implies that you can learn how to take a particular test method, and that there can be method effect. Thus method effect and washback appear to be linked indirectly if not directly.

It is interesting to note that scores on the New Exam decline rapidly from Type A schools through to Type D schools, and for Types C and D, the scores are very low indeed: Khaniya concludes that these students have learnt very little English indeed, despite having higher SLC scores. (There are many students in his sample who pass SLC but get miserably low New Exam scores.) Thus he claims that SLC does not measure English ability, but something else. However, SLC and the new exam correlate at .72. Interestingly, this varies by school type as follows: A= -.12 (!)
B= .73
C= .62
D= .63

Interestingly also, the Cambridge O-Level test correlates at -0.06 with the New Exam and .15 with SLC, and the relationship between college teacher ratings of students and SLC is .67, whereas with the New Exam it is .07!

Khaniya (1990b) also administered questionnaires to SLC teachers, and 50% claimed that they were free to teach what they thought would benefit their students in Grade 10 (ie,

they were not obliged by the exam to do particular things. Yet 67% claim they have to spend a lot of time preparing students for the SLC exam. 80% said however that they had to prepare answers to possible questions on the exam, and 75% confessed they did "question spotting". However, we are given no details on what actually happened in classes, rather than what teachers said they did.

Studies of the clasroom impact of examinations are very rare indeed. The only study identified to date is the Sri Lankan O-Level Evaluation Project, which is the subject of a separate paper (see Wall and Alderson, this symposium).

Our tentative conclusions are that the impact at least of the Sri Lankan O-Level is less pervasive than we had expected, and we are currently trying to understand why this might be the case. As Wall and Alderson suggest, it may be because of the teachers' lack of information about the examination. It may also be because of lack of understanding on the part of teachers of what might be an appropriate way to prepare students for the examination. It may even be because the exam itself - and this may indeed be true of all exams - does not and cannot <u>determine</u> how teachers teach, however much it might <u>influence</u> what they teach. This has important consequences for the nature of the Washback Hypothesis.

A series of proposals for research

Clearly more research is needed in this area. We have already suggested that it is important to define what is meant by the term washback: what scope it should have, and where its limits lie, and what aspects of impact we do not wish to include in the concept of washback. Secondly, it is important to state explicitly what one's version of the Washback Hypothesis is: it is highly likely that it will be more complex than the fifteen hypotheses. Nevertheless, it will be necessary to spell out in some detail what the predicted effects of the test are, and it is quite likely that this statement will have to take account of the nature of the test concerned, the educational context within which it is used and the nature of the decisions that are taken on the basis of the test results.

In addition, in parallel to this increasing specification of the hypothesis, it will be important to take account of findings in the research literature in at least two areas: that of motivation and performance, and that of innovation and change in educational settings.

Then we will need to consider the methodology to be used in research into washback. There has been a tendency to date to relay upon participants' reported perceptions of events through questionnaire responses, or to examine the results and relationships of test performances. Smith (1991) is instructive with respect to possible methods:

BEST COPY AVAILABLE

> "In (Smith et al, 1989) we employed direct observation of classrooms, meetings and school life generally; interviews with teachers, pupils, administrators, and others; and analysis of documents... To understand the perceived effects of external testing on teachers, one only needs to ask. Their statements on questionnaires, in interviews, and during conversations in meetings and lounges reveal the anxiety, shame, loss of esteem and alienation they experience from publication and use of test scores."

We suggest, however, that it is increasingly obvious that we need to look closely at classroom events in particular, in order to see whether what teachers and learners say they do is reflected in their behaviour.

In addition, we believe it important in conjunction with classroom observations to triangulate the researcher's perceptions of events with some account from participants of how they perceived and reacted to events in class, as well as outside - this amounts to an advocacy of a more ethnographic approach to the topic than has been common heretofor (see Watson-Gegeo, 1988 for a clear discussion of this issue).

Finally, as well as attempting to describe the washback that occurs, we need to attempt, at some point in the future, to account for what occurs, and this is likely to involve widening our hypothesis formulation and data collection to include explanatory variables derived from the research literature mentioned above.

What this amounts to is a long-term and relatively complex research program. Given the considerations we adduce in this paper, we believe this is both inevitable as well as desirable. What is undesirable is a continuation of our state of ignorance about a phenomenon on whose importance all seem to be agreed. Equaly undesirable is a continuation of naive assertions about washback on the part of applied linguists in general, materials writers, syllabus designers, teachers, as well as language testers, until some empirical investigations have been undertaken!

Bibliography

Alderson, J Charles (1986) "Innovations in Language Testing?" in Portal (ed)

Alderson, J Charles, Clapham, C and Wall, D (1987) An Evaluation of the National Certificate in English, 1986 Institute for English Language Education, Lancaster University.

Alderson, J Charles and North, B (eds) (1991) Language Testing in the 1990s: The Communicative Legacy. Developments in Language Teaching series, Modern English Publications.

Alderson, J Charles and Wall, D (1990) The Sri Lankan O-Level Evaluation Project: Second Interim Report. Lancaster University

Alderson J Charles and Wall, D (1991) The Sri Lankan O-Level Evaluation Project: Third Interim Report. Lancaster University

Chamberlain, D and Baumgardner, R (eds) (1988) ESP in the Classroom: Practice and Evaluation. ELT Document 128. Modern English Publications

Davies, A. (1968) Language Testing Symposium: A Psycholinguistic Approach. Oxford University Press

Davies, A. (1985) "Follow My Leader: Is That What Language Tests Do?" in Lee et al.

Davies, A, Glendinning E H and McLean A C (1984) The English Language Teaching Survey of Nepal, 1983-84

Ebel R L (1979) Essentials of Educational Measurement Third edition Prentice-Hall inc.

Forbes, D (1973) "Selling English Short". English Language Teaching Journal, XXVII, p132-137.

Fransson, A (1984) "Cramming or Understanding? Effects of intrinsic and extrinsic motivation on approach to learning and test performance" in Alderson, J Charles and A H Urquhart (eds) Reading in a Foreign Language: London: Longman, pp86 - 115

Frederiksen, N (1984) "The Real Test Bias" American Psychologist, March pp 193- 202

Frederiksen, J R and A Collins (1989) "A Systems Approach to Educational Testing". Educational Researcher,Vol 18, no 9, pp 27-32

Haladyna, T H, S B Nolen and N S Haas (1991) "Raising Standardized Achievement Test Scores and the Origins of Test Score Pollution" in Educational Researcher, Vol20 No 5 pp 2- 7

Hughes, A (1988) "Introducing a Needs-based test of English Language Proficiency into an English-medium University in Turkey" in Hughes, A (ed) pp134-153

Hughes, A (ed) (1988) Testing English for University Study ELT Documents 127 Modern English Publications

Hughes, A (1989) Testing for Language Teachers Cambridge University Press

Kellaghan, T, Madaus, G F and Airasian, P W (1982) The Effects of Standardized Testing. London: Kluwen, Nijholf Publishing

Khaniya, T R (1990a) "The Washback Effect of a Textbook-Based Test". Edinburgh Working Papers in Applied Linguistics Number 1: pp 48-58

Khaniya, T R (1990b) "Examinations as Instruments for Educational Change: Investigating the Washback Effect of the Nepalese English Exams". Unpublished PhD thesis, University of Edinburgh, Scotland.

Lauwers and Seanlon (eds) (REM?? see Khaniya and Wong, below??)

Lee, et al (1985) New Directions in Language Testing Pergamon Press

McDonough, S (1981) Psychology in Foreign Language Teaching. Hemel Hempstead: Allen and Unwin.

Madsen, H (1976) "New Alternatives in EFL Exams or 'How to Avoid Selling English Short'" English Language Teaching Journal, Vol XXX No 2 pp135-144

Morris, B (1972) Objectives and Perspectives in Education: Studies in Educational Theories. London. Routledge and Kegan Paul

Morrow, K (1986) "The Evaluation of Tests of Communicative Performance" in Portal (ed)

Paris, S G, T A Lawton, J C Turner and J L Roth (1991) "A Developmental Perspective on Standardized Achevement Testing" Educational Researcher, Vol 20, No 5 pp 12 - 19

Pearson, I (1988) "Tests as Levers for Change" in Chamberlain and Baumgardner (eds) pp 98 - 107

Pilliner, A (1973) "Assessment - Principles and Practice with Special Reference to Education in Pakistan". Unpublished ms, The British Council

Portal, M (ed) (1986) Innovations in Language Testing. NFER/Nelson.

Smith, M L (1991) Put to the Test: The Effects of External Testing on Teachers. Educational Researcher, Vol 20, No 5, pp 8-11

Swain, M (1985) REM??Large-scale Communicative Testing"?? in Lee et al

Wall, D (1991) "Measuring Examination 'Washback': The Sri Lankan Evaluation Project". Paper presented at the IAEA Conference, Nairobi, Kenya, May

Wall, D, Clapham, C and Alderson, J Charles (1991) "Validating Tests in Difficult Circumstances" in Alderson and North (eds) pp209-225

Watson-Gegeo, K A (1988) "Ethnography in ESL: Defining the Essentials". TESOL QUARTERLY, Vol 22 No 4, pp 575 - 592.

Wesdorp, Hildo (1982) Backwash effects of language testing in primary and secondary education. Stichting Centrum voor onderwijsonderzoek van de Universiteit van Amsterdam, Postbus 3753, 1001 AN Amsterdam.

Wiseman, S (ed) (1961) Examinations and English Education. Manchester University Press

Wong R H K (1969) Educational Effects of Examinations on Pupils, Teachers and Society in Lauwerys and Seanlon (eds)